ADAPTIVE THRESHOLDS FOR NEURAL NETWORKS WITH SYNAPTIC NOISE

D. BOLLÉ and R. HEYLEN
Institute for Theoretical Physics, Katholieke Universiteit Leuven
Celestijnenlaan 200 D, B-3001, Leuven, Belgium
E-mail: desire.bolle@fys.kuleuven.be / rob.heylen@fys.kuleuven.be

The inclusion of a macroscopic adaptive threshold is studied for the retrieval dynamics of both layered feedforward and fully connected neural network models with synaptic noise. These two types of architectures require a different method to be solved numerically. In both cases it is shown that, if the threshold is chosen appropriately as a function of the cross-talk noise and of the activity of the stored patterns, adapting itself automatically in the course of the recall process, an autonomous functioning of the network is guaranteed. This self-control mechanism considerably improves the quality of retrieval, in particular the storage capacity, the basins of attraction and the mutual information content.

1. Introduction

In general pattern recognition problems, information is mostly encoded by a small fraction of bits and also in neurophysiological studies the activity level of real neurons is found to be low, such that any reasonable network model has to allow variable activity of the neurons. The limit of low activity, i.e., sparse coding is then especially interesting. Indeed, sparsely coded models have a very large storage capacity behaving as $1/(a \ln a)$ for small a, where a is the activity (see, e.g., [1, 2, 3, 4] and references therein). However, for low activity the basins of attraction might become very small and the information content in a single pattern is reduced [4]. Therefore, the necessity for a control of the activity of the neurons has been emphasized such that the latter stays the same as the activity of the stored patterns during the recall process. This has led to several discussions imposing external constraints on the dynamics of the network. However, the enforcement of such a constraint at every time step destroys part of the autonomous functioning of the network, i.e., a functioning that has to be independent precisely from such external constraints or control mechanisms. To solve this problem, quite recently a self-control mechanism has been introduced in the dynamics of networks for so-called diluted architectures [5]. This self-control mechanism introduces a time-dependent threshold in the transfer function [5, 6]. It is determined as a function of both the cross-talk noise and the activity of the stored patterns in the network, and adapts itself in the course of the recall process. It furthermore allows to reach optimal retrieval performance both in the absence and in the presence of synaptic noise [5, 6, 7, 8]. These diluted architectures contain no common ancestors nodes, in contrast with feedforward architectures. It has then been shown that a similar mechanism can be introduced succesfully for layered feedforward architectures but, without synaptic noise [9]. Also for fully connected neural networks, the idea of self-control has been partially exploited for three-state neurons [10]. However, due to the feedback correlations present in such an architecture, the dynamics had to be solved approximately and again, without synaptic noise.

The purpose of the present work is twofold: to generalise this self-control mechanism for layered architectures when synaptic noise is allowed, and to extend the idea of self-control in fully connected networks with exact dynamics and synaptic noise. In both cases it can be shown that it leads to a substantial improvement of the quality of retrieval, in particular the storage capacity, the basins of attraction and the mutual information content.

The rest of the paper is organized as follows. In Sections 2 and 3 the layered network is treated. The precise formulation of the layered model is given in Section 2 and the adaptive threshold dynamics is studied in Section 3. In Sections 4 and 5 the fully connected network is studied. The model set-up and its exact threshold dynamics is described in Section 4, the numerical treatment and results are presented in Section 5. Finally, Section 6 contains the conclusions.

2. The layered model

Consider a neural network composed of binary neurons arranged in layers, each layer containing N neurons. A neuron can take values $\sigma_i(t) \in \{0,1\}$ where $t = 1,\ldots,L$ is the layer index and $i = 1,\ldots,N$ labels the neurons. Each neuron on layer t is unidirectionally connected to all neurons on layer $t+1$. We want to memorize p patterns $\{\xi_i^\mu(t)\}, i = 1,\ldots,N,\ \mu = 1,\ldots,p$ on each layer t, taking the values $\{0,1\}$. They are assumed to be independent identically distributed random variables with respect to i, μ and t, determined by the probability distribution

$$p(\xi_i^\mu(t)) = a\delta(\xi_i^\mu(t)-1) + (1-a)\delta(\xi_i^\mu(t)) \qquad (1)$$

From this form we find that the expectation value and the variance of the patterns are given by $E[\xi_i^\mu(t)] = E[\xi_i^\mu(t)^2] = a$. Moreover, no statistical correlations occur, in fact for $\mu \neq \nu$ the covariance vanishes.

The state $\sigma_i(t+1)$ of neuron i on layer $t+1$ is determined by the state of the neurons on the previous layer t according to the stochastic rule

$$P(\sigma_i(t+1) \mid \boldsymbol{\sigma}(t)) = \frac{1}{1+e^{2(2\sigma_i(t+1)-1)\beta h_i(t)}}. \qquad (2)$$

with $\boldsymbol{\sigma}(t) = (\sigma_1(t), \sigma_2(t), \ldots, \sigma_N(t))$. The right hand side is the logistic function. The "temperature" $T = 1/\beta$ controls the stochasticity of the network dynamics, it measures the synaptic noise level [11]. Given the network state $\boldsymbol{\sigma}(t)$ on layer t, the so-called "local field" $h_i(t)$ of neuron i on the next layer $t+1$ is given by

$$h_i(t) = \sum_{j=1}^{N} J_{ij}(t)(\sigma_j(t)-a) - \theta(t) \qquad (3)$$

with $\theta(t)$ the threshold to be specified later. The couplings $J_{ij}(t)$ are the synaptic strengths of the interaction between neuron j on layer t and neuron i on layer $t+1$. They depend on the stored patterns at different layers according to the covariance rule

$$J_{ij}(t) = \frac{1}{Na(1-a)} \sum_{\mu=1}^{N} (\xi_i^\mu(t+1)-a)(\xi_j^\mu(t)-a)\ . \qquad (4)$$

These couplings then permit to store sets of patterns to be retrieved by the layered network.

The dynamics of this network is defined as follows (see [12]). Initially the first layer (the input) is externally set in some fixed state. In response to that, all neurons of the second layer update synchronously at the next time step, according to the stochastic rule (2), and so on.

At this point we remark that the couplings (4) are of infinite range (each neuron interacts with infinitely many others) such that our model allows a so-called mean-field theory approximation. This essentially means that we focus on the dynamics of a single neuron while replacing all the other neurons by an average background local field. In other words, no fluctuations of the other neurons are taken into account. In our case this approximation becomes exact because, crudely speaking, $h_i(t)$ is the sum of very many terms and a central limit theorem can be applied [11].

It is standard knowledge by now that mean-field theory dynamics can be solved exactly for these layered architectures (e.g., [12, 13]). By exact analytic treatment we mean that, given the state of the first layer as initial state, the state on layer t that results from the dynamics is predicted by recursion formulas. This is essentially due to the fact that the representations of the patterns on different layers are chosen independently. Hence, the big advantage is that this will allow us to determine the effects from self-control in an exact way.

The relevant parameters describing the solution of this dynamics are the *main overlap* of the state of the network and the μ-th pattern, and the *neural activity* of the neurons

$$M^\mu(t) = \frac{1}{Na(1-a)} \sum_{i=1}^{N} (\xi_i^\mu(t)-a)(\sigma_i(t)-a) \qquad (5)$$

$$q(t) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i(t)\ . \qquad (6)$$

In order to measure the retrieval quality of the recall process, we use the mutual information function [5, 6, 14, 15]. In general, it measures the average amount of information that can be received by the user by observing the signal at the output of a channel [16, 17]. For the recall process of stored patterns that we are discussing here, at each layer the process can be regarded as a channel with input $\xi_i^\mu(t)$ and output $\sigma_i(t)$ such that this mutual information function can be defined as [5, 16]

$$I(\sigma_i(t);\xi_i^\mu(t)) = S(\sigma_i(t)) - \langle S(\sigma_i(t)|\xi_i^\mu(t))\rangle_{\xi^\mu(t)} \tag{7}$$

where $S(\sigma_i(t))$ and $S(\sigma_i(t)|\xi_i^\mu(t))$ are the entropy and the conditional entropy of the output, respectively

$$\begin{aligned} S(\sigma_i(t)) &= -\sum_{\sigma_i} p(\sigma_i(t))\ln[p(\sigma_i(t))] \quad (8) \\ S(\sigma_i(t)|\xi_i^\mu(t)) &= -\sum_{\sigma_i} p(\sigma_i(t)|\xi_i^\mu(t)) \\ &\quad \times \ln[p(\sigma_i(t)|\xi_i^\mu(t))]\ . \quad (9) \end{aligned}$$

These information entropies are peculiar to the probability distributions of the output. The quantity $p(\sigma_i(t))$ denotes the probability distribution for the neurons at layer t and $p(\sigma_i(t)|\xi_i^\mu(t))$ indicates the conditional probability that the i-th neuron is in a state $\sigma_i(t)$ at layer t given that the i-th site of the pattern to be retrieved is $\xi_i^\mu(t)$. Hereby, we have assumed that the conditional probability of all the neurons factorizes, i.e., $p(\{\sigma_i(t)\}|\{\xi_i(t)\}) = \prod_j p(\sigma_j(t)|\xi_j(t))$, which is a consequence of the mean-field theory character of our model explained above. We remark that a similar factorization has also been used in Schwenker et al. [18].

The calculation of the different terms in the expression (7) proceeds as follows. Because of the mean-field character of our model the following formulas hold for every neuron i on each layer t. Formally writing (forgetting about the pattern index μ) $\langle O\rangle \equiv \langle\langle O\rangle_{\sigma|\xi}\rangle_\xi = \sum_\xi p(\xi)\sum_\sigma p(\sigma|\xi)O$ for an arbitrary quantity O the conditional probability can be obtained in a rather straightforward way by using the complete knowledge about the system: $\langle\xi\rangle = a$, $\langle\sigma\rangle = q$, $\langle(\sigma-a)(\xi-a)\rangle = M$, $\langle 1\rangle = 1$.

The result reads

$$\begin{aligned} p(\sigma|\xi) &= [\gamma_0 + (\gamma_1-\gamma_0)\xi]\ \delta(\sigma-1) \\ &\quad +[1-\gamma_0-(\gamma_1-\gamma_0)\xi]\ \delta(\sigma) \quad (10) \end{aligned}$$

where $\gamma_0 = q - aM$ and $\gamma_1 = (1-a)M + q$, and where the M and q are precisely the relevant parameters (5) for large N. Using the probability distribution of the patterns we obtain

$$p(\sigma) = q\delta(\sigma-1) + (1-q)\delta(\sigma)\ . \tag{11}$$

Hence the entropy (8) and the conditional entropy (9) become

$$\begin{aligned} S(\sigma) &= -\ q\ln q - (1-q)\ln(1-q) \quad (12) \\ S(\sigma|\xi) &= -\ [\gamma_0 + (\gamma_1-\gamma_0)\xi]\ln[\gamma_0 + (\gamma_1-\gamma_0)\xi] \\ &\quad -\ [1-\gamma_0-(\gamma_1-\gamma_0)\xi] \\ &\qquad \times\ln[1-\gamma_0-(\gamma_1-\gamma_0)\xi]\ . \quad (13) \end{aligned}$$

By averaging the conditional entropy over the pattern ξ we finally get for the mutual information function (7) for the layered model

$$\begin{aligned} I(\sigma;\xi) &= -q\ln q - (1-q)\ln(1-q) \\ &+\ a[\gamma_1\ln\gamma_1 + (1-\gamma_1)\ln(1-\gamma_1)] \\ &+\ (1-a)[\gamma_0\ln\gamma_0 + (1-\gamma_0)\ln(1-\gamma_0)]\ . \end{aligned} \tag{14}$$

3. Adaptive thresholds in the layered network

It is standard knowledge (e.g., [12]) that the synchronous dynamics for layered architectures can be solved exactly following the method based upon a signal-to-noise analysis of the local field (3) (e.g., [4, 13, 19, 20] and references therein). Without loss of generality we focus on the recall of one pattern, say $\mu = 1$, meaning that only $M^1(t)$ is macroscopic, i.e., of order 1 and the rest of the patterns causes a cross-talk noise at each step of the dynamics.

We suppose that the initial state of the network model $\{\sigma_i(1)\}$ is a collection of independent identically distributed random variables, with average and variance given by $E[\sigma_i(1)] = E[(\sigma_i(1))^2] = q_0$. We furthermore assume that this state is correlated with only one stored pattern, say pattern $\mu = 1$, such that $\mathrm{Cov}(\xi_i^\mu(1),\sigma_i(1)) = \delta_{\mu,1}\ M_0^1\ a(1-a)$.

Then the full recall proces is described by [12, 13]

$$M^1(t+1) = \frac{1}{2}\int \mathcal{D}x \left(\tanh\left[\beta F_1\right] + \tanh\left[\beta F_2\right]\right) \tag{15}$$

$$q(t+1) = aM^1(t+1) + \frac{1}{2}\left(1 + \int \mathcal{D}x \tanh\left[\beta F_2\right]\right) \tag{16}$$

$$D(t+1) = Q(t+1) + \frac{\beta}{2}\left\{1 - a\int \mathcal{D}x \tanh^2[\beta F_1] - (1-a)\int \mathcal{D}x \tanh^2[\beta F_2]\right\}^2 D(t) \tag{17}$$

with

$$F_1 = (1-a)M^1(t) - \theta(t) + \sqrt{\alpha D(t)}\,x \tag{18}$$

$$F_2 = -aM^1(t) - \theta(t) + \sqrt{\alpha D(t)}\,x \tag{19}$$

and $\alpha = p/N$, $\mathcal{D}x$ is the Gaussian measure $\mathcal{D}x = dx(2\pi)^{-1/2}\exp(-x^2/2)$, where $Q(t) = [(1-2a)q(t) + a^2]$ and where $D(t)$ contains the influence of the cross-talk noise caused by the patterns $\mu > 1$. As mentioned before, $\theta(t)$ is an adaptive threshold that has to be chosen.

In the sequel we discuss two different choices and both will be compared for networks with synaptic noise and various activities. Of course, it is known that the quality of the recall process is influenced by the cross-talk noise. An idea is then to introduce a threshold that adapts itself autonomously in the course of the recall process and that counters, at each layer, the cross-talk noise. This is the self-control method proposed in [5]. This has been studied for layered neural network models without synaptic noise, i.e., at $T = 0$, where the rule (2) reduces to the deterministic form $\sigma_i(t+1) = \Theta(h_i(t))$ with $\Theta(x)$ the Heaviside function taking the value $\{0, 1\}$. For sparsely coded models, meaning that the pattern activity a is very small and tends to zero for N large, it has been found [9] that

$$\theta(t)_{sc} = c(a)\sqrt{\alpha D(t)}, \quad c(a) = \sqrt{-2\ln a} \tag{20}$$

makes the second term on the r.h.s of Eq.(16) at $T = 0$, asymptotically vanish faster than a such that $q \sim a$. It turns out that the inclusion of this self-control threshold considerably improves the quality of retrieval, in particular the storage capacity, the basins of attraction and the information content.

The second approach chooses a threshold by maximizing the information content, $i = \alpha I$ of the network (recall Eq. (14)). This function depends on $M^1(t)$, $q(t)$, a, α and β. The evolution of $M^1(t)$ and of $q(t)$ (15), (16) depends on the specific choice of the threshold through the local field (3). We consider a layer independent threshold $\theta(t) = \theta$ and calculate the value of (14) for fixed a, α, M_0^1, q_0 and β. The optimal threshold, $\theta = \theta_{opt}$, is then the one for which the mutual information function is maximal. The latter is non-trivial because it is even rather difficult, especially in the limit of sparse coding, to choose a threshold interval by hand such that i is non-zero. The computational cost will thus be larger compared to the one of the self-control approach. To illustrate this we plot in Fig. 1 the information content i as a function of θ without self-control or a priori optimization, for $a = 0.005$ and different values of α. For every value of α, below its critical value, there is a range for the threshold where the information content is different from zero and hence, retrieval is possible. This retrieval range becomes very small when the storage capacity approaches its critical value $\alpha_c = 6.4$.

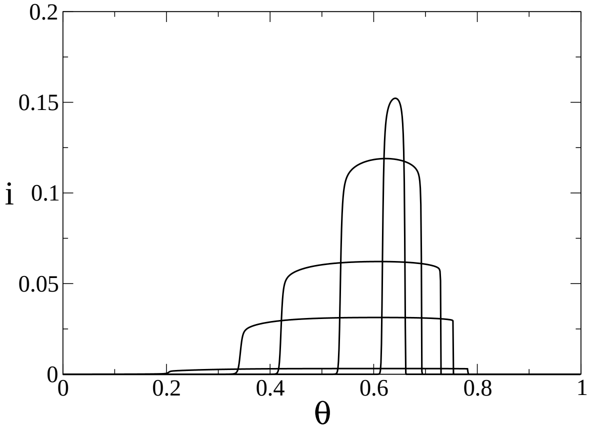


Figure 1: The information $i = \alpha I$ as a function of θ for $a = 0.005$, $T = 0.1$ and several values of the load parameter $\alpha = 0.1, 1, 2, 4, 6$ (bottom to top)

Concerning then the self-control approach, the next problem to be posed in analogy with the case without synaptic noise is the following one. Can one determine a form for the threshold $\theta(t)$ such that the integral in the second term on the r.h.s of Eq.(16) at $T \neq 0$ vanishes asymptotically faster than a?

In contrast with the case at zero temperature where due to the simple form of the transfer function, this threshold could be determined analytically (recall Eq. (20)), a detailed study of the asymptotics of the integral in Eq. (16) gives no satisfactory analytic solution. Therefore, we have designed a systematic numerical procedure through the following steps:

- Choose a small value for the activity a'.
- Determine through numerical integration the threshold θ' such that

$$\int_{-\infty}^{\infty} \frac{dx\ e^{-x^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \Theta(x-\theta) \le a' \quad \text{for} \quad \theta > \theta' \tag{21}$$

  for different values of the variance $\sigma^2 = \alpha D(t)$.

- Determine as a function of $T = 1/\beta$, the value for θ'_T such that for $\theta > \theta' + \theta'_T$

$$\int_{-\infty}^{\infty} \frac{dx\ e^{-y^2/\sigma^2}}{2\sigma\sqrt{2\pi}} [1 + \tanh[\beta(x-\theta)]] \le a' \tag{22}$$

The second step leads precisely to a threshold having the form of Eq. (20). The third step determining the temperature-dependent part θ'_T leads to the final proposal

$$\theta_t(a,T) = \sqrt{-2\ln(a)\alpha D(t)} - \frac{1}{2}\ln(a)T^2. \tag{23}$$

This dynamical threshold is again a macroscopic parameter, thus no average must be taken over the microscopic random variables at each step t of the recall process.

We have solved these self-controlled dynamics, Eqs.(15)-(17) and (23), for our model with synaptic noise, in the limit of sparse coding, numerically. In particular, we have studied in detail the influence of the T-dependent part of the threshold. Of course, we are only interested in the retrieval solutions with $M > 0$ (we forget about the index 1) and carrying a non-zero information $i = \alpha I$. The important features of the solution are illustrated, for a typical value of a in Figs. 2-4. In Fig. 2 we show the basin of attraction for the whole retrieval phase for the model with threshold (20) (dashed curves) compared to the model with the noise-dependent threshold (23) (full curves). We see that there is no clear improvement for low T but there is a substantial one for higher T. Even near the border of critical storage the results are still improved such that also the storage capacity itself is larger.

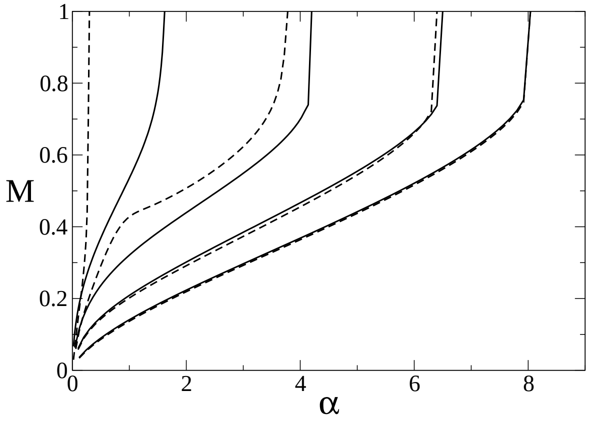


Figure 2: The basin of attraction as a function of α for $a = 0.005$ and $T = 0.2, 0.15, 0.1, 0.05$ (from left to right) with (full lines) and without (dashed lines) the T-dependent part in the threshold (23).

This is further illustrated in Fig. 3 where we compare the evolution of the retrieval overlap $M(t)$ starting from several initial values, M_0, for the model without (Fig. 3 (a)) and with (Fig. 3 (b)) the T-correction in the threshold and for the optimal threshold model (Fig. 3 (c)). Here this temperature correction is absolutely crucial to guarantee retrieval, i.e., $M \approx 1$. It really makes the difference between retrieval and non-retrieval in the model. Furthermore, the model with the self-control threshold with noise-correction has even a

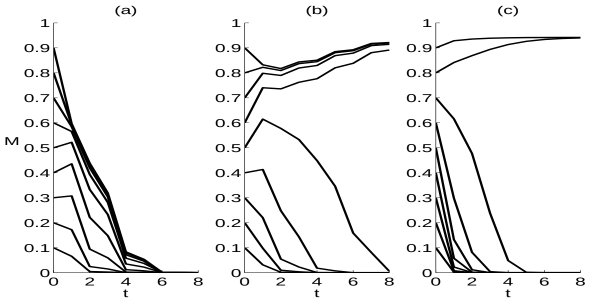


Figure 3: The evolution of the main overlap $M(t)$ for several initial values M_0 with $T = 0.2$, $q_0 = a = 0.005$, $\alpha = 1$ for the self-control model (23) without (a) and with T-dependent part (b) and for the optimal threshold model (c).

wider basin of attraction than the model with optimal threshold.

In Fig. 4 we plot the information content i as a function of the temperature for the self-control dynamics with the threshold (23) (full curves), respectively (20) (dashed curves). We see that a substantial improvement of the information content is obtained.

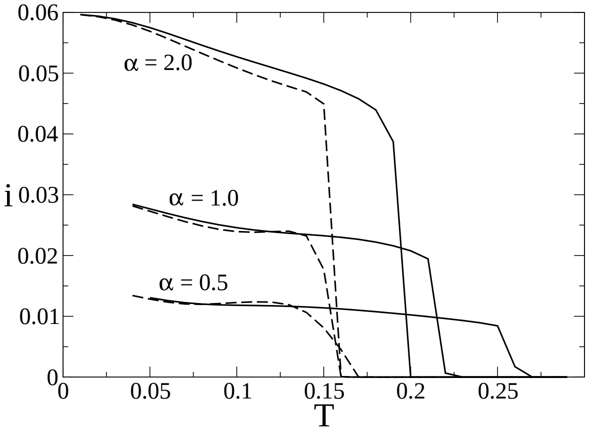
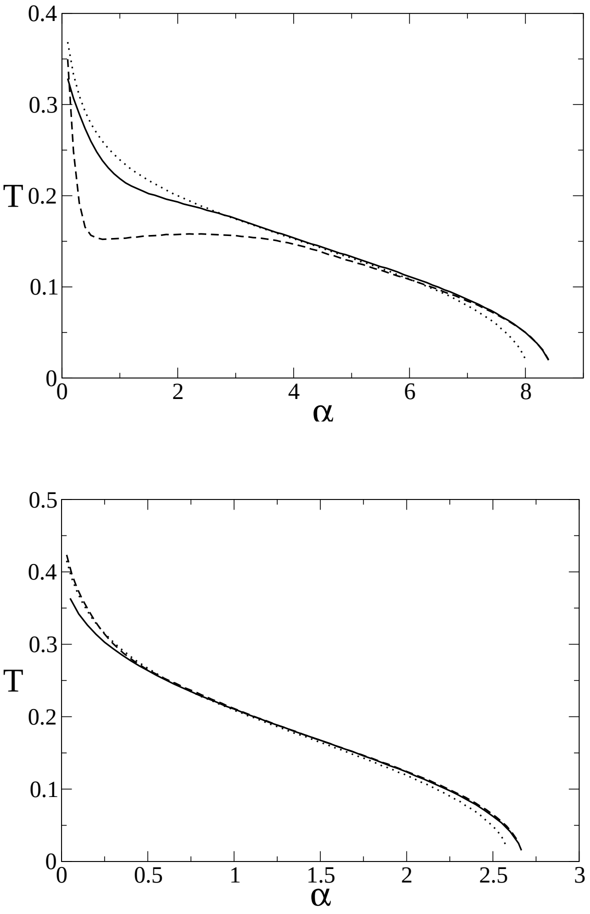


Figure 4: The information content $i = \alpha I$ as a function of T for several values of the loading α and $a = 0.005$ with (full lines) and without (dashed lines) the T-correction in the threshold.

Finally we show in Fig. 5 a $T - \alpha$ plot for $a = 0.005$ (a) and $a = 0.02$ (b) with (full line) and without (dashed line) noise-correction in the self-control threshold and with optimal threshold (dotted line). These lines indicate two phases of the layered model: below the lines our model allows recall, above the lines it does not. For $a = 0.005$ we see that the T-dependent term in the self-control threshold leads to a big improvement in the region for large noise and small loading and in the region of critical loading. For $a = 0.02$ the results for the self-control threshold with and without noise-correction and those for the optimal thresholds almost coincide, but we recall that the calculation with self-control is autonomously done by the network and less demanding computationally.

In the next Sections we want to find out whether this self-control mechanism also works in the fully connected network for which we work out the dynamics in the presence of synaptic noise in an exact way. We start by defining the model and describing this dynamics.



(b)

Figure 5: Phases in the $T - \alpha$ plane for $a = 0.005$ (a) and $a = 0.02$ (b) with (full line) and without (dashed line) the temperature correction in the self-control threshold and with optimal threshold (dotted line).

4. Dynamics of the fully connected model

As before, the network we consider consists of N binary neurons $\sigma_i \in \{0, 1\}, i = 1 \ldots N$ but the couplings J_{ij} between each pair of neurons σ_i and σ_j are now given by the following rule

$$J_{ij} = \sum_{\mu=1}^{p} (\xi_i^\mu - a)(\xi_j^\mu - a) \qquad (24)$$

The local field is now determined by

$$h_i(\boldsymbol{\sigma}, t) = \frac{1}{a(1-a)N} \sum_{j=1}^{N} J_{ij}\sigma_j(t) + \theta(\boldsymbol{q}) \qquad (25)$$

The threshold is represented by the function θ and, based upon the results obtained in the previous sections and in [10] we have chosen this to be a function of the mean activity q of the neurons.

In order to study the dynamics of this model we need to define the transition probabilities for going from one state of the network to another. For each neuron at time $t+1$, $\sigma_i(t+1)$, we have the following stochastic rule (compare (2))

$$P(\sigma_i(t+1)|\boldsymbol{\sigma}(t)) = \frac{\exp(-\beta\epsilon(\sigma_i(t+1)|\boldsymbol{\sigma}(t))}{\sum_s \exp(-\beta\epsilon(s|\boldsymbol{\sigma}(t))} \qquad (26)$$

where

$$\epsilon(\sigma_i(t+1)|\boldsymbol{\sigma}(t)) = -\sigma_i(t+1)h_i(\boldsymbol{\sigma}(t)) \qquad (27)$$

with the local fields given by (25) and where $\boldsymbol{\sigma}(0)$ at time $t = 0$ is the known starting configuration.

The dynamics is then described using the generating function analysis, which was introduced in [21] to the field of statistical mechanics and, by now, is part of many textbooks. The idea of this approach to study dynamics [21, 22] is to look at the probability to find a certain microscopic path in time. The basic tool to study the statistics of these paths is the generating functional

$$Z[\boldsymbol{\psi}] = \left\langle \sum_{\boldsymbol{\sigma}(0)\ldots\boldsymbol{\sigma}(t)} P(\boldsymbol{\sigma}(0), \ldots, \boldsymbol{\sigma}(t)) e^{-i\sum_i \sum_{s=1}^{t} \psi_i(s)\sigma_i(s)} \right\rangle_{\boldsymbol{\xi}} \qquad (28)$$

with $P(\boldsymbol{\sigma}(0), \ldots, \boldsymbol{\sigma}(t))$ the probability to have a certain path in phase space

$$\begin{aligned} & P(\boldsymbol{\sigma}(0), \ldots, \boldsymbol{\sigma}(t)) \\ &\quad = P(\boldsymbol{\sigma}(0)) \prod_{s=1}^{t} W[\boldsymbol{\sigma}(s-1), \boldsymbol{\sigma}(s)] && (29) \\ &\quad = P(\boldsymbol{\sigma}(0)) \prod_{s=1}^{t} \prod_{i=1}^{N} P(\sigma_i(s)|\boldsymbol{\sigma}(s-1)) && (30) \end{aligned}$$

Here $W[\boldsymbol{\sigma}, \boldsymbol{\tau}]$ is the transition probability for going from the configuration $\boldsymbol{\sigma}$ to the configuration $\boldsymbol{\tau}$, and the $P(\sigma_i(s)|\boldsymbol{\sigma}(s-1))$ are given by (26). In (28) the average over the patterns $\boldsymbol{\xi}$ has to be taken since they are independent identically distributed random variables, determined by the probability distribution (1).

One can find all physical observables by including a time-independent external field $\gamma_i(t)$ in (27) in order to define a response fuction, and then calculating appropriate derivatives of (28) with respect to $\psi_i(s)$ or $\gamma_i(t)$ letting all $\psi_i(t); i = 1, \ldots, N$ tend to zero afterwards. For example we can write the main overlap $m(s)$ (as before we focus on the recall of one pattern), the correlation function $C(s, s')$ and the response function $G(s, s')$ as

$$\begin{aligned} m(s) &= \frac{1}{a(1-a)N} \sum_i \xi_i \sigma_i(s) \\ &= i \lim_{\boldsymbol{\psi}\to 0} \frac{1}{a(1-a)N} \sum_i \xi_i \frac{\delta Z}{\delta\psi_i(s)} && (31) \\ C(s, s') &= \frac{1}{N} \sum_i \sigma_i(s)\sigma_i(s') \\ &= -\lim_{\boldsymbol{\psi}\to 0} \frac{1}{N} \sum_i \frac{\delta^2 Z}{\delta\psi_i(s)\delta\psi_i(s')} && (32) \\ G(s, s') &= \frac{1}{N} \sum_i \frac{\delta}{\delta\gamma_i(s')} \sigma_i(s) \\ &= i \lim_{\boldsymbol{\psi}\to 0} \frac{1}{N} \sum_i \frac{\delta^2 Z}{\delta\psi_i(s)\delta\gamma_i(s')} && (33) \end{aligned}$$

The further calculation is rather technical, and we point the interested reader to the literature for more details (e.g.,[22, 23]). One obtains an effective single neuron local field given by

$$\begin{aligned} h(s) &= \frac{1}{a(1-a)} \left(\boldsymbol{m}(s) - a\boldsymbol{q}(s)\right)(\xi - a) + \theta(\boldsymbol{q}) \\ &\quad + \alpha \sum_{s'=0}^{s-1} R(s, s')\sigma(s') + \sqrt{\alpha}\eta(s) && (34) \end{aligned}$$

with $\eta(s)$ temporally correlated noise with zero mean and correlation matrix $\boldsymbol{D}$, and the retarded self-interaction $\boldsymbol{R}$ which are given by

$$\begin{aligned} \boldsymbol{D} &= (\boldsymbol{1} - \boldsymbol{G})^{-1} \boldsymbol{C} (\boldsymbol{1} - \boldsymbol{G}^{\dagger})^{-1} && (35) \\ \boldsymbol{R} &= (\boldsymbol{1} - \boldsymbol{G})^{-1} && (36) \end{aligned}$$

The final result for the evolution equations of the physical observables is given by four self-consistent equations

$$\begin{aligned} m(s) &= \langle \xi\sigma(s)\rangle_* && (37)\\ q(s) &= \langle \sigma(s)\rangle_* && (38)\\ C(s,s') &= \langle \sigma(s)\sigma(s')\rangle_* && (39)\\ G(s,s') &= \beta\Big\langle \sigma(s)\Big[\sigma(s'+1)- \\ &\qquad \Big(1+e^{\beta h(\boldsymbol{\sigma},\boldsymbol{\eta},s')}\Big)^{-1}\Big]\Big\rangle_* && (40)\end{aligned}$$

The average over the effective path measure and the recalled pattern $\langle\cdot\rangle_*$ is given by

$$\langle g\rangle_* = \sum_{\xi} p(\xi) \sum_{\sigma(0),\ldots,\sigma(t)} \int d\boldsymbol{\eta} P(\boldsymbol{\eta}) P(\boldsymbol{\sigma} \mid \boldsymbol{\eta}) g \qquad (41)$$

with $p(\xi)$ given by (1), $d\boldsymbol{\eta} = \prod_{s'} d\eta(s')$ and with

$$\begin{aligned} P(\boldsymbol{\eta}) &= \frac{1}{\sqrt{\det(2\pi\boldsymbol{D})}} \\ &\times\ \exp\left(-\frac{1}{2}\sum_{s,s'=0}^{t-1} \eta(s)\boldsymbol{D}^{-1}(s,s')\eta(s')\right) && (42)\\ P(\boldsymbol{\sigma} \mid \boldsymbol{\eta}) &= (1+m(0)(2\sigma(0)-1)-\sigma(0)) \\ &\qquad \times \left(\prod_{s=1}^{t} \frac{e^{\beta\sigma(s)h(s-1)}}{1+e^{\beta h(s-1)}}\right) && (43)\end{aligned}$$

Remark that the term involving the one-time observables in (34) has the form $(\boldsymbol{m} - a\boldsymbol{q})$. Therefore, in the sequel we define the main overlap M as

$$M = \frac{1}{a(1-a)}(\boldsymbol{m} - a\boldsymbol{q}) \quad \in [-1,1] \qquad (44)$$

The set of equations (37), (38), (39) and (40) represent an exact dynamical scheme for the evolution of the network.

To solve these equations numerically we use the Eisfeller and Opper method ([24]). The algorithm these authors propose is an advanced Monte-Carlo algorithm. Recalling equation (41) this requires samples from the correlated noise (for the integrals over η), the neurons (for the sums) and the pattern variable ξ. Instead of generating the complete vectors at each timestep, we represent these samples by a large population of individual paths, where each path consists of t neuron values, t noise values and one pattern variable. All the averages (integrations, sums and traces over probability distributions) can then be represented by summations over this population of single neuron evolutions. Because of causality, we also know that it is possible to calculate a neuron at time s when we know all the variables (neurons, noise, physical observables) at previous timesteps. Also, the initial configuration at time zero is known. This gives rise to an iterative scheme allowing us to numerically solve the equations at hand.

The main idea then is to represent the average (41) over the statistics of the single particle problem, as an average over the population of single neuron evolutions. Since we did not find an explicit algorithm in the literature we think that it is very useful to write one down explicitly.

- Choose a large number K, the number of independent neuron evolutions in the population, a final time t_f, an activity a, a pattern loading α, and an initial condition (an initial overlap, correlation, activity, ...).

- Generate space for K neuron evolutions p_i. Each evolution contains a pattern variable $\xi_i \in \{0,1\}$, t_f neuron variables $\sigma_i(s) \in \{0,1\}$, and t_f noise variables $\eta_i(s) \in \mathbb{R}, s = 0\ldots t_f, i = 1\ldots K$.

- At time 0, initialize the ξ_i according to the distribution (1). Then initialize the neuron variables at time zero employing the initial condition, e.g.:

  When an initial activity is defined:
  $$P(\sigma_i(0)=1) = q(0)$$
  When an initial overlap is defined:
  $$P(\sigma_i(0)=\xi_i) = M(0)$$

- The algorithm is recursive. So, at time t we assume that we know the neuron variables for all times $s \leq t$, the noise variables for all times $s < t$, and the matrix elements $D(s,s')$ for $s,s' < t$. We want to first calculate the noise variables at time t, and then the neuron variables at time $t+1$. At timestep t this can be done as follows

  1. Calculate the physical observables $m(t)$, $q(t)$ and $C(t,s) = C(s,t)$, $s \leq t$, by sum-

ming over the population:

$$m(t) = \frac{1}{K}\sum_{i=1}^{K} \xi_i \sigma_i(t) \quad (45)$$

$$q(t) = \frac{1}{K}\sum_{i=1}^{K} \sigma_i(t) \quad (46)$$

$$C(t,s) = \frac{1}{K}\sum_{i=1}^{K} \sigma_i(t)\sigma_i(s) \quad (47)$$

2. For $s < t$ calculate the matrix $\boldsymbol{L}$

$$L(t,s) = \frac{1}{K}\sum_{i=1}^{K} \sigma_i(t)\eta_i(s) \quad (48)$$

3. Calculate $\boldsymbol{G} = \alpha^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1}$, where $\boldsymbol{D}$ is the known noise correlation matrix from the previous timestep. Turn $\boldsymbol{G}$ into a square matrix by adding a column of zeros to the end.

4. Calculate $\boldsymbol{R} = (\boldsymbol{1} - \boldsymbol{G})^{-1}$ and the new $\boldsymbol{D} = \boldsymbol{R}\boldsymbol{C}\boldsymbol{R}^{\dagger}$

5. For each site i, calculate a new noise variable:

$$\eta_i(t) = \frac{\zeta_i(t)}{\sqrt{\boldsymbol{D}^{-1}(t,t)}} - \frac{1}{\boldsymbol{D}^{-1}(t,t)}\sum_{s<t} \boldsymbol{D}^{-1}(t,s)\eta_i(s) \quad (49)$$

   where all $\zeta_i(t)$ are independently chosen from a standard gaussian distribution.

6. Calculate the effective local field at each site:

$$h_i(t) = M(t)\,(\xi_i - a) + \theta(q(t)) + \alpha\sum_{s\leq t} R(t,s)\sigma_i(s) + \sqrt{\alpha}\eta_i(t) \quad (50)$$

7. Use this local field to determine the new spin value at each site at time $t+1$:

$$P(\sigma_i(t+1)) = \frac{e^{\beta\sigma_i(t+1)h_i(t)}}{1+e^{\beta h_i(t)}} \quad (51)$$

8. If $t < t_f$ increase t and go to step 1. Else stop.

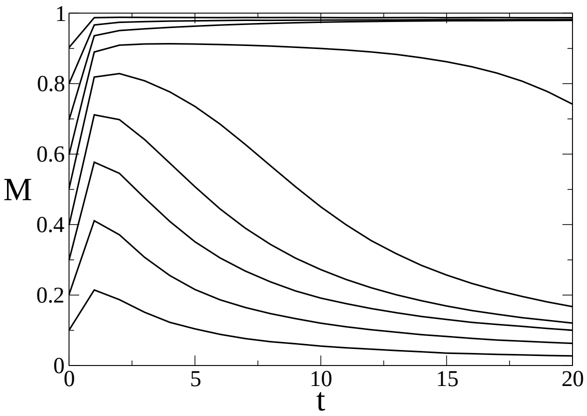


Figure 6: The evolution of the overlap of the fully connected network for several initial overlaps. The system parameters are $\alpha = 0.06$, $a = 0.5$, $T = 0.04$ and $\theta(q) = 0$.

This algorithm can be easily performed in a parallel way. All individual neuron evolutions are independent of each other, and the only steps that cannot be executed in a distributed fashion are steps 3 and 4. It turns out that these two steps mostly take less than 1% of the total calculation time.

5. Thresholds in the fully connected network

We have used the algorithm above to check the evolution of the overlap. The threshold function $\theta(q(t))$ appears in the local field (50), and its effect on the evolution of the different physical observables can be investigated.

We take the size of the population of independent neuron evolutions $K = 10^6$. Larger population sizes can be obtained by making the algorithm parallel, but no significant differences are found.

We first look at the unbiased case ($a = 1/2$) without threshold. In fig. 6 we plot the evolution of the overlap M for several initial conditions. When the initial overlap M_0 is too smal there is no retrieval. This critical initial overlap separating a retrieval phase from a non-retrieval phase forms the border of the basin of attraction. For biased low activity networks, it is already known (e.g, [4]) that a constant threshold ($a-1/2$) has to be introduced in the local field eq. (25) in order to guarantee a correct functioning of the network. This can easily be seen by noting that for a network where only one single

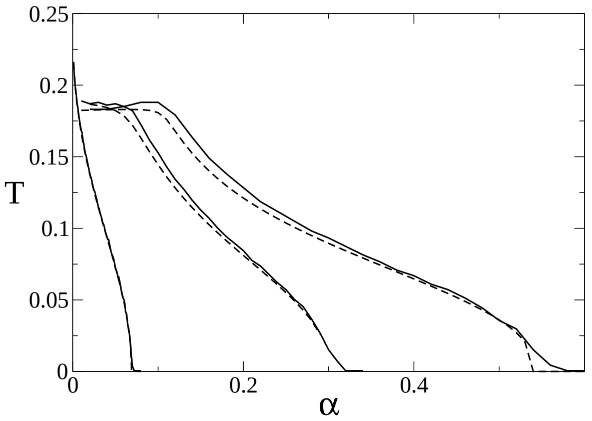


Figure 7: Phases in the $T-\alpha$ plane for, from left to right, $a=0.5$, $a=0.1$, $a=0.05$ with $\theta(q)=a-0.5$. Solid (dashed) lines indicate the results for the dynamics (statics).

pattern is stored ($h_i \to \xi_i - a$) such that the field becomes $(1-a)$ or $(-a)$. And this lies completely asymmetric with respect to the symmetric (around the point 1/2) transfer function eq.(26). For $a \to 0$ one even finds that the probability that a neuron changes its state from zero to one becomes 1/2.

A $T-\alpha$ plot for several values of the activity with $\theta(q)=a-0.5$ is presented in fig. 7. The solid lines represent the results from the dynamics obtained by initializing the algorithm discussed in section 4 with an initial overlap $M_0=1$, and determining the temperature where this overlap has decreased below 0.4 after 200 timesteps. For comparison the dashed lines show the results from an equilibrium statistical mechanics calculation (e.g., [25, 26]). As to be expected, both calculations agree. These lines indicate two phases of the fully connected model: below the lines our model allows recall, above the lines it does not.

The main question we want to address in this Section is whether we can again improve the retrieval capacities of this network architecture by introducing the self-control threshold (23). We recall that the quantity $D(t)$ occurring in this expression contains the influence of the cross-talk noise. From the signal-to-noise ratio analysis in [10] and from statistical neurodynamics arguments ([20]) we know that the leading term of $D(t)$ is $q(t)$. Moreover, from a biological point of view, it does not seem plausible that a network monitors the statistical quantity of

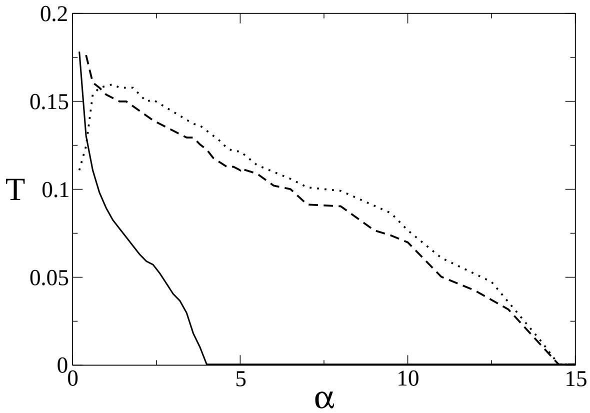


Figure 8: Phases in the $T-\alpha$ plane for $a=0.005$ and several thresholds. Solid: $\theta=a-0.5$; dashed: self-control threshold without T-correction; dotted: self-control threshold with T-correction.

the cross-talk noise. Therefore, we take $D(t)=q(t)$ in the self-control threshold in fully connected networks.

We have then solved the generating functional analysis (37)-(40) with the threshold

$$\theta(q(t)) = \sqrt{-2\ln(a)\alpha q(t)} - \frac{1}{2}\ln(a)T^2. \qquad (52)$$

Some typical results are shown in figs. 8-10. For system parameters comparable with those for the layered architecture, fig. 8 clearly shows that the self-control threshold without T-correction significantly increases the retrieval region, and the temperature correction further improves the results for α not too small.

Looking at a fixed $T=0.1$ for this case (Fig. 9), we furthermore notice that the self-control threshold without T-correction again significantly increases the basin of attraction. The additional temperature correction further increases this basin, and even increases the maximal achievable pattern loading α.

For lower temperatures (Fig. 10) the self-control threshold still increases the basin of attraction for larger values of the pattern loading α, but for smaller loadings the effect is diminishing. The temperature correction gives no clear improvement in this case. A similar behavior was observed for the layered architecture in fig. 2. We remark that the

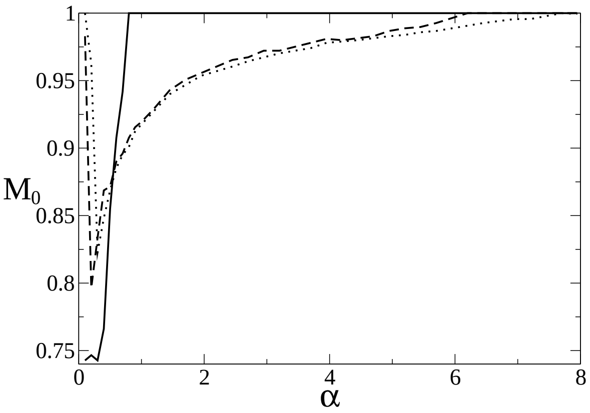


Figure 9: The basin of attraction as a function of α for $a = 0.005$ and $T = 0.1$. Solid: $\theta = a - 0.5$; dashed: self-control threshold without T-correction; dotted: self-control threshold with T-correction.

subtraction of $(a - 1/2)$ is not necessary when using the self-control method. The latter takes this into account automatically and the networks operates fully autonomously.

6. Conclusions

In this work we have studied the inclusion of an adaptive threshold in sparsely coded layered and fully connected neural networks with synaptic noise. We have presented an analytic form for a self-control threshold, allowing an autonomous functioning of these networks, and compared it, for the layered architecture, with an optimal threshold obtained by maximizing the mutual information which has to be calculated externally each time one of the network parameters (activity, loading, temperature) is changed. The consequences of this self-control mechanism on the quality of the recall process have been studied.

We find that the basins of attraction of the retrieval solutions as well as the storage capacity are enlarged. For some activities the self-control threshold even sets the border between retrieval and non-retrieval. This confirms the considerable improvement of the quality of recall by self-control, also for layered and fully connected network models with synaptic noise.

This allows us to conjecture that self-control might be relevant even for dynamical systems in general, when trying to improve, e.g., basins of attraction.

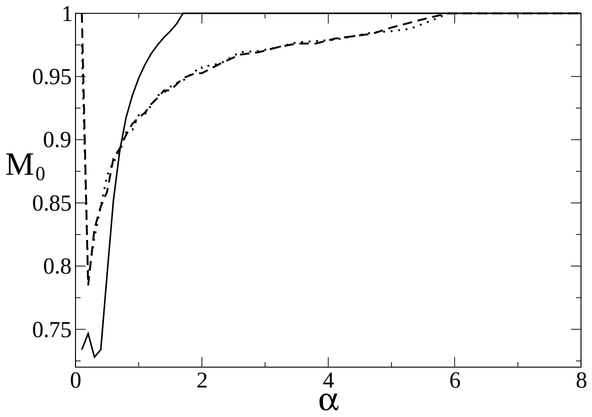


Figure 10: The basin of attraction as a function of α for $a = 0.01$ and $T = 0.05$. Solid: constant $\theta = a - 0.5$; dashed: self-control threshold without T-correction; dotted: self-control threshold with T-correction.

Acknowledgment

This work has been supported by the Fund for Scientific Research- Flanders (Belgium).

References

1. D. J. Willshaw, O. P. Buneman and H. C. Longuet-Higgins 1969, "Nonholographic associative memory“, *Nature* **222**, 960.
2. G. Palm 1981, "On the storage capacity of an associative memory with random distributed storage elements“, *Biol. Cyber.* **39**, 125.
3. E. Gardner 1988, "The space of interactions in neural network models“, *J. Phys. A: Math. Gen.* **21**, 257.
4. M. Okada 1996, "Notions of associative memory and sparse coding“, *Neural Networks* **9**, 1429.
5. D. R. C. Dominguez and D. Bollé 1998, "Self-control in sparsely coded networks“, *Phys. Rev. Lett.* **80**, 2961.
6. D. Bollé, D. R. C. Dominguez and S. Amari 2000, "Mutual information of sparsely coded associative memory with self-control and ternary neurons“, *Neural Networks* **13**, 455.
7. D. Bollé and R. Heylen 2004, "Self-control dynamics for sparsely coded networks with synaptic noise“, in *2004 Proceedings of the IEEE International Joint Conference on Neural Networks*, p.3195
8. D. R. C. Dominguez, E. Korutcheva, W. K.

Theumann and R. Erichsen Jr. 2002, "Flow diagrams of the quadratic neural network", *Lecture Notes in Computer Science*, **2415**, 129.

9. D. Bollé G. and Massolo 2000, "Thresholds in layered neural networks with variable activity", *J. Phys. A: Math. Gen.* **33**, 2597.
10. D. Bollé and D. R. C. Dominguez 2000, "Mutual information and self-control of a fully-connected low-activity neural network", *Physica A* **286**, 401.
11. J. Hertz, A. Krogh and R. G. Palmer 1991, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City.
12. E. Domany, W. Kinzel and R. Meir 1989, "Layered Neural Networks", *J.Phys. A: Math. Gen.* **22**, 2081.
13. D. Bollé 2004, "Multi-state neural networks based upon spin-glasses: a biased overview", in *Advances in Condensed Matter and Statistical Mechanics* eds. E. Korutcheva and R. Cuerno, Nova Science Publishers, New-York, p. 321-349.
14. J-P. Nadal, N. Brunel and N. Parga 1998, "Nonlinear feedforward networks with stochastic outputs: infomax implies redundancy reduction", *Network: Computation in Neural Systems* **9**, 207.
15. S. Schultz and A. Treves 1998, "Stability of the replica-symmetric solution for the information conveyed by a neural network", *Phys. Rev. E* **57**, 3302.
16. R. E. Blahut 1990, *Principles and Practice of Information Theory*, Reading, MA: Addison-Wesley.
17. C. E. Shannon 1948, "A mathematical theory for communication", *Bell Systems Technical Journal* **27**, 379.
18. F. Schwenker, F. T. Sommer and G. Palm 1996, "Iterative retrieval of sparsely coded associative memory patterns", *Neural Networks* **9**, 445.
19. S. Amari 1977, "Neural theory and association of concept information", *Biol. Cyber.* **26**, 175.
20. S. Amari and K. Maginu 1988, "Statistical neurodynamics of associative memory", *Neural Networks* **1**, 63.
21. P. C. Martin, E. D. Siggia and H. A. Rose 1973, "Statistical dynamics of classical systems", *Phys. Rev. A* **8**, 423.
22. A. C. C. Coolen 2001, "Statistical mechanics of recurrent neural networks II: dynamics", in *Handbook of biological physics*, vol. 4, eds. F. Moss and S. Gielen, Elsevier Science.
23. D. Bollé, J. Busquets Blanco and T. Verbeiren 2004, "The signal-to-noise analysis of the Little-Hopfield model revisited", *J. Phys. A: Math. Gen.* **37**, 1951.
24. H. Eissfeller and M. Opper 1992, "New method for studying the dynamics of disordered spin systems without finite-size effects", *Phys. Rev. Lett.* **68**, 2094.
25. H. Horner 1989, "Neural networks with low levels of activity: Ising vs. McCulloch-Pitts neurons", *Z. Phys. B. - Cond. Mat.* **75**, 133.
26. D. J. Amit, H. Gutfruend and H. Sompolinsky 1987, "Information storage in neural networks with low levels of activity", *Phys. Rev. A* **35**, 2293.